Matemaattisen mallinnuksen soveltamisesta populaatiogenetiikassa

Pro Gradu-tutkielma
Markus Tapani Hyytinen
2373162
Matemaattisten tieteiden yksikkö
Oulun yliopisto
29.3.2019

Sisältö

Johdanto

"Maailma on jakautunut tosiasioihin. . . . Me teemme itsellemme kuvia tosiasioista. . . . Kuvan looginen muoto on ilmentää sitä, mitä se kuvaa. . . . Kuvalla on kuvattavan kanssa yhteistä itse kuvaamisen looginen muoto. Kuva kuvastaa todellisuutta esittämällä mahdollisuuden tosiasioiden olemassaoloon tai -olemattomuuteen. . . . Looginen kuva tosiasioiden olemassaolosta tai -olemattomuudesta on väitelause. . . . Matematiikka on looginen metodi. Matematiikan väitelauseet ovat yhtälöitä. . . . Jos kysymys on mahdollista muotoilla, voidaan siihen myös vastata."

Lainaus on Ludwig Wittgensteinin teoksen *Tractatus Logico-Philosophicus* osioista 1.2, 2.1, 2.2, 2.201, 4.1, 6.2 ja 6.5 [38].

Mallinnettaessa luonnonilmiötä matemaattisesti, maailman tosiasiallinen olotila puetaan loogisen kielen muotoon. Tällöin tosiasioista voidaan tehdä loogisia päätelmiä siten, että päättelyä itsessään ei voida kyseenalaistaa: seuraukset ovat välttämättömiä lähtötilanteesta, jolloin lopputulosta voidaan pitää eräänlaisena tautologiana lähtötilanteesta [38]. Heikommassa muodossa voidaan väittää, että päättelyn lopputulos on yhtä luotettava, kuin sen lähtötilanne oli. Looginen päättely on lisäksi yksiselitteistä ja toistettavaa, ja on siten helposti jaettavissa tiedeyhteisön keskustelun piiriin.

Täydellistä varmuutta saavutetaan kuitenkin hyvin harvoin, jollei milloinkaan. Albert Einstein on sanonut:

"Miten voi olla, että matematiikka kuvastaa todellisuuden olioita niin ihailtavasti, vaikka se on vain ihmisajattelun tulos ja lisäksi irrallinen kokemuksesta? . . . Mielestäni lyhyt vastaus tähän kysymykseen on, että kun matemaattiset lait viittaavat todellisuuteen ne ovat epävarmoja ja kun ne ovat varmoja, ne eivät viittaa todellisuuteen."[26]

Ja Bertrand Russell puolestaan:

"Fysiikka ei ole matemaattista siksi, että tietäisimme niin paljon fyysisestä maailmasta, vaan siksi, että tiedämme kovin vähän: voimme löytää vain sen matemaattisia ominaisuuksia."[29]

Matemaattiseen mallintamiseen liittyy siis olennaisesti epävarmuus: Vastaako malli todellisuutta? Voidaanko mallin perusteella tehdä ennustuksia, jotka toteutuvat vaaditulla todennäköisyydellä? Voidaanko tarkasteltavasta asianlaidasta tai ajatuksesta edes tehdä mallia, siis puhua siitä matematiikan kielellä? Wittgensteinin mukaan asioista, joita ei voi muotoilla logiikan kielellä, ei pitäisi edes puhua logiikkaa hyödyntäen [38, osio 7]. Juuri tätä kuitenkin tehdään, kun laaditaan tilastollisia malleja biologisista prosesseista. Todennäköisyyslaskenta ja tilastotiede antaa

työkaluja hallita epävarmuutta empiirisissä mittaustuloksissa. Lisäksi biologisia prosesseja pyritään palauttamaan biokemiallisiksi prosesseiksi ja vielä kemiallisiksi prosesseiksi virhelähteiden löytämiseksi ja eristämiseksi. Matemaattiset tulokset kuten keskeinen raja-arvolause ja suurten lukujen laki, modernin tilastotieteen perustavia lauseita, mahdollistavat empiirisistä tuloksista puhumisen *hyödyllisellä* tavalla, vaikka täydellistä varmuutta ei saavutettaisiinkaan.

Tässä työssä esitellään johdantoa matemaattisen mallintamisen filosofiaan ja parametrin käsitteen soveltamiseen malleja laadittaessa. Populaatioiden mallintamista käsitellään esimerkkien kautta ja populaatiogenetiikasta esitellään kaksi keskeistä tulosta: Hardyn-Weinbergin laki, joka on nykyään laajalti käytetty työkalu, mutta historiallisesti huomattava merkkipaalu todennäköisyyslaskennan käyttöönotosta metodina biologian sisällä. Toinen esiteltävä tulos on Tajiman D-koe, moderni tilastollinen koe, jota käytetään mm. luonnonvalinnan havaitsemiseen geneettisestä aineistosta.

1 Mallintaminen

Józef Baranyi erottaa *empiirisen mallin* ja *matemaattisen mallin* käsitteet:

"Puhtaasti empiirinen malli ... on myös malli, mutta sen tarkoituksena ei ole muuta kuin esittää empiiriset mittaustulokset elegantisti. Se on (*regressio*) malli, siinä mielessä kuin sanaa käytetään tilastotieteessä, kun tavoitteena on erilaisten vasteiden esittäminen numeerisesti yksinkertaisten funktioiden, kuten polynomien avulla ilman mekanistista selitystä. Termi *matemaattinen* malli on huolellisemmin määritelty ja viittaa joukkoon perustavanlaatuisia hypoteesejä tutkittaviin prosesseihin liittyen, joista osa saattaa ilmentyä funktioina ja (differentiaali-) yhtälöinä. Siispä mekanistisesta näkökulmasta *funktio* ja *malli* eivät ole sama asia. Funktio on matemaattinen abstraktio, joka helpottaa tietyn mallin kuvailua."[1] (Oma suomennos, alkuperäinen kursivointi.)

Itse ajatustyön vaihteittaista rakennetta mallintamisprojektissa voidaan kuvata seuraavasti:

1. Ilmiö
2. Matemaattinen malli
3. Mittausten suunnittelu
4. Mittaustulokset
5. Mallin parametrien estimointi

6. Mallin parantaminen, takaisin vaiheeseen 2
7. Mallin soveltaminen käytäntöön

Seuraavassa luvussa esitellään yleisimpiä mallintamisfunktioita populaatiomallintamisen näkökulmasta.

1.1 Yleisiä mallintamisfunktioita

Populaatioiden mallinnus differentiaaliyhtälöillä on mielekästä, vaikka elossa olevien yksilöiden määrä onkin diskreetti, aina positiivinen muuttuja. Kun populaatio on tarpeeksi suuri, on yhden yksilön muutos hyvin pieni verrattuna koko populaation kokoon tehden pyöristysvirheistä merkityksettömiä. Myöhemmin huomataan, että mallinnettavan populaation koko on merkittävä suure myös muilla tavoin. Yksinkertaisimmillaan eliöpopulaation elossaolevien yksilöiden määrää ajan funktiona voidaan kuvata yhden parametrin avulla. Olkoon $p(t)$ populaation koko ajanhetkellä t ja $r(t,p)$ populaation syntyvyyden ja kuolleisuuden erotus aikayksikössä. Jos populaatio on eristyksissä muista populaatioista eli isolaatiossa, niin muuttoliikkeen vaikutus parametriin $r(t,p)$ on olematon. On myös mahdollista sisällyttää muuttoliike parametriin $r(t,p)$ seuraavasti:

$$r(t,p) = \text{syntyvyys} + \text{saapuvat yksilöt} - \text{kuolleisuus} - \text{lähtevät yksilöt}$$

Nyt voidaan lausua populaation yksilömäärän muutos dp/dt yksinkertaisesti:

$$dp/dt = r(t,p) \cdot p(t)$$

Kaikkein yksinkertaisimmassa mallissa oletetaan, että r on vakio, eli se pysyy samana kaiken aikaa minkä tahansa suuruisissa populaatioissa. Tästä oletuksesta seuraa differentiaaliyhtälö, jota kutsutaan nimellä *lineaarinen yhtälö* tai *Malthusin laki* (Malthusian Law of Population Growth tai vain eksponentiaalisen kasvun malli)

$$\frac{dp(t)}{dt} = ap(t), \quad a = \text{vakio} \tag{1}$$

Tästä differentiaaliyhtälöstä voidaan ratkaista funktio $p(t)$ jos tiedetään yksilömäärä p_0 jollain ajanhetkellä t_0. Tällöin alkuarvo-ongelman

$$\frac{dp(t)}{dt} = ap(t), \quad p(t_0) = p_0$$

ratkaisu on $p(t) = p_0 e^{a(t-t_0)}$. Tästä huomataan, että Malthusin lakia noudattavat populaatiot kasvavat eksponentiaalisesti, jos kasvavat ollenkaan.

Joissakin tapauksissa tämä yksinkertainen lineaarinen malli on riittävä populaation kasvun mallintamiseen. Esimerkiksi kolibakteeripopulaation kasvua elintarviketurvallisuuden laboratoriotutkimuksissa on mallinnettu onnistuneesti yksinkertaisella eksponentiaalisella kasvumallilla, ja huomattu sen antavan yhtä hyvän sovitteen kuin "monimutkaisemmat" mallit [5]. On kuitenkin huomattu, että kun populaatio kasvaa hyvin suureksi elinoloihinsa nähden, malli lakkaa toimimasta. Tämä on ymmärrettävää, sillä malli ei ota huomioon elinympäristön epätäydellisyyttä ja eliöiden kilpailua resursseista. Lineaariseen malliin on siis lisättävä negatiivinen termi, joka kuvastaa kilpailua yksilöiden välillä. Pitäydytään populaation sisäisessä kilpailussa, sillä lajin sisällä yksilöt kilpailevat resursseista saman ekolokeron sisällä. Sopiva valinta kilpailutermiksi on $-bp^2$, jossa b on populaatiolle ominainen vakio ja p^2 kuvastaa populaation yksilöiden välisten kohtaamisten keskimääräistä lukumäärää aikayksikössä. Mitä paremmin resurssit riittävät populaatiossa, sitä vähemmän yksilöiden välinen kilpailu vaikuttaa selviytymiseen, aiheuttaen termille b pienemmän arvon, kuin jos resurssit olisivat niukat. Muokatuksi yhtälöksi saadaan:

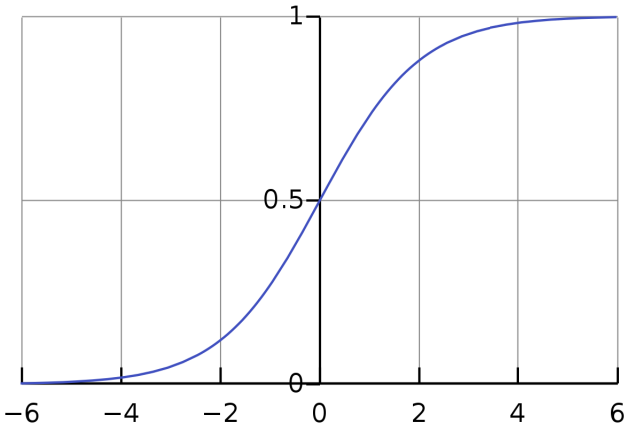
$$\frac{dp(t)}{dt} = ap - bp^2. \tag{2}$$

Tämä yhtälö on *Logistinen populaatiokasvun malli* (Logistic Law of Population Growth). Parametreja a ja b kutsutaan tässä mallissa populaation elinvoimaisuuskertoimiksi (vital coefficient). Tässä mallissa populaation kasvu on ensin lähes eksponentiaalista hidastuen myöhemmin. Logistinen malli liittyy läheisesti *logistiseen funktioon*, jota logistisen mallin kehittäjä Pierre François Verhulst tutki:

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}}$$

Funktiossa L on käyrän eli populaation maksimiarvo, x_0 käyrän symmetriakeskipiste ja k käyrän jyrkkyys. Logistisen funktion maksimiarvo voitaisiin helposti tulkita suurimmaksi mahdolliseksi populaation kooksi, jota ympäristö pystyy ylläpitämään. Tämä päättely ei kuitenkaan ole täysin pätevää, sillä termi $-bp^2$ valittiin kuvastamaan kahden yksilön suorasta tai epäsuorasta kohtaamisesta johtuvaa kilpailua ottamatta kantaa populaation ulkoisen ympäristön vaikutukseen.

Logistisen mallin vakio b on yleensä paljon pienempi kuin vakio a. Tästä seuraa, että pienillä muuttujan p arvoilla termi $-bp^2$ on merkityksetön verrattuna termiin ap, jolloin logistinen malli palautuu lineaariseen malliin ja populaatio kasvaa eksponentiaalisesti. Kun taas muuttuja p on suuri, vaikuttaa termi $-bp^2$ merkittävästi populaation kasvun hidastamiseen. Esimerkiksi teollisuusvaltioiden väkilukua mallinnettaessa mitä enemmän elintilaa ja ruokaa väestöllä on, sitä pienempi



Kuva 1: Logistisen funktion perusmuoto

parametri b on. Pienillä populaatioilla malli voi törmätä ongelmiin, kun todellisuuden satunnaishäiriöiden, kuten onnettomuuksien, painoarvo kasvaa. Suurilla populaatioilla suurten lukujen laki ja keskeinen raja-arvolause silottavat satunnaisilmiöiden merkitystä.

Esimerkki lineaarisesta mallista löytyy kappaleesta 2.1.

1.2 Parametrien esimointi

Tehtäessä kvantitatiivista tutkimusta luonnonilmiöistä tavoitteena on yleensä saada tilastollista tietoa ilmiöiden, tapahtumien, ympäristön ja olioiden olemassaolosta tai niiden välisistä syy-seuraussuhteista. Tutkimukseen liittyy usein matemaattinen malli tai malleja, joiden perusteella tutkimuksen mittaukset suunnitellaan. Usein tutkittavaa ilmiötä tai ilmiöiden suhdetta ei voi mitata suoraan, jolloin mallin perusteella pitää määritellä mielekkäästi mitattava muuttuja. Toisinaan lähestyminen on päinvastainen: käytettävissä olevat mittausvälineet ovat rajallisia tai mittaus on jo suoritettu, jolloin mittaustulosten tulkintaan ja tutkimiseen käytetään erilaisia tilastollisia menetelmiä ja vertaillaan niiden pohjalta tehtyjen päätelmien mielekkyyttä. Parametrien estimointi on suoraan yhteydessä sekä mittaustuloksiin että käytettyyn matemaattiseen malliin ja soveltuu kumpaankin tutkimisen lähestymistapaan.

Parametrien estimoinnissa mittaustulokset nähdään yleisen matemaattisen mallin erityistapauksena. Mittaustulokset antavat tilastollista tukea mallin parametrien arvoiksi tai vaihteluväleiksi. Eri malleja voidaan vertailla laskemalla, kuinka vahvaa tukea mittaustulokset antavat niille. Mallien vertailua käsitellään tarkemmin

kappaleessa 1.4. Yleisten matemaattisten mallien käyttö on tärkeää tutkimustulosten tulkinnan kannalta. Laajalti käytettyjen mallien parametrit ovat helposti globaalin tiedeyhteisön ymmärrettävissä ja tulkittavissa. Yksinkertaisesti tulkittavat parametrit ovat myös muidenkin kuin matemaatikkojen ymmärrettävissä, esimerkiksi syntyvyys, kuolleisuus tai jonkin ominaisuuden esiintymistiheys populaatiossa tietyllä ajanhetkellä.

Estimoitaessa mallin parametreja mittaustuloksista arvioidaan todennäköisyyttä, että todellinen maailma on matemaattisen mallin mukainen. Tämä todennäköisyys sisältää epävarmuutta, joka jakautuu epätasaisesti mallin parametrien välille. Mittausprosessin eri virheet aiheuttavat epävarmuutta, kuten myös ilmiön todellinen yhteensopimattomuus matemaattiseen malliin joko kokonaan tai osittain. Myös tutkimuskysymyksen muotoilu aiheuttaa epävarmuutta.

Samat mittaustulokset voivat antaa tukea usealle eri tutkimuskysymykselle samasta ilmiöstä ja myös eri vastauksille, toisinaan jopa vastakkaisille. Tutkijan tehtävänä parametrien estimoinnissa on selvittää, millaista ja kuinka voimakasta tukea mittaustulokset antavat eri tutkimuskysymyksille ja tulkinnoille. Tässä tutkija tarvitsee erityisalansa asiantuntemusta, mutta yhtä lailla tärkeää on itse todennäköisyyden käsitteen ymmärtäminen ja tulkitseminen, jotta tutkijalla on mahdollisimman laaja keskusteluavaruus tulkintojen keksimistä ja tarkastelua varten.

1.3 Todennäköisyyden tulkitsemisen eri koulukunnat

Tämän alaluvun päälähteenä on käytetty sivuston *Stanford Encyclopedia of Philosophy* artikkelia todennäköisyyden käsitteen historiasta ja tulkinnasta [13]. Todennäköisyyden käsitteeseen sisältyy usein intuitiivisia käsityksiä, joiden tiedostaminen ja selittäminen on tärkeä osa tilastollisen tutkimuksen tietoteoreettista pohjaa. Alan Hájek listaa kolme tärkeintä konseptia, mitä todennäköisyyteen yleensä liitetään [13]:

1. Epätarkka looginen konsepti, joka mittaa todistusaineiston antamaa tukea tulkinnalle objektiivisesti, esimerkiksi: "Seismologisten mittaustulosten mukaan Kaliforniassa *todennäköisesti* tapahtuu suuri maanjäristys tällä vuosikymmenellä".

2. Konsepti ajattelijan luottamuksesta uskomukseensa jonkinlaisella määrittelemättömällä asteikolla, esimerkiksi "En ole varma että Samarkandissa sataa tällä viikolla, mutta *todennäköisesti* ei".

3. Objektiivinen konsepti, joka pätee maailmanlaajuisesti eri järjestelmiin riippumatta ajattelijoiden mielipiteistä, esimerkiksi: "Mielivaltaisesti valittu radium-

ydin hajoaa *todennäköisesti* 10000 vuodessa".

1.3.1 Klassinen todennäköisyys

Klassisen koulukunnan todennäköisyyslaskentaa kehittivät mm. Laplace ja Pascal. Klassisessa tulkinnassa todennäköisyys jaetaan tasan mahdollisten alkeistapausten välille. Tähän tulkintaan liittyy implisiittisesti oletus todistusaineiston puutteesta tai täydellisestä symmetriasta. Lisäksi alkeistapausten oletetaan olevan "samantyyppisiä", mikä käy ilmi Laplacen todennäköisyyksiä käsittelevästä esseestä:

"The theory of chance consists in reducing all the events of the same kind to a certain number of cases equally possible, that is to say, to such as we may be equally undecided about in regard to their existence, and in determining the number of cases favorable to the event whose probability is sought. The ratio of this number to that of all the cases possible is the measure of this probability, which is thus simply a fraction whose numerator is the number of favorable cases and whose denominator is the number of all the cases possible."[21]

Päädytään siis kaavaan:

$$p = \frac{\text{suotuisten alkeistapausten lukumäärä}}{\text{kaikkien alkeistapausten lukumäärä}}$$

Klassista tulkintaa voidaan kritisoida esimerkiksi väitteellä, että oletusta todistusaineiston puutteesta olisi parempi mallintaa epätarkalla todennäköisyydellä kuten $p \in [0,1]$ tai jättämällä todennäköisyysarvio kokonaan esittämättä. Tässä kritiikin kohteena on siis klassiseen tulkintaan sisältyvä oletus, että paremman tiedon puutteessa alkeistapausten esiintyminen on yhtä todennäköistä. Esimerkiksi kolikonheitossa kruunan ja klaavan todennäköisyys oletetaan samaksi, mutta esimerkiksi Belgian vuoden 2001 yhden euron ja pienempien kolikkojen väitettiin tuottavan heitettäessä hieman todennäköisemmin kruunan kuin klaavan, vaikka tämä väite myöhemmin kumottiinkin [22].

Lisäksi alkeistapausten "samantyyppisyys"ei ole yksiselitteistä. Kolikonheitossa on harvinaista, mutta mahdollista, että kolikko tippuu reunalleen eikä kummallekaan tyypilliselle sivulle. Myös odottamattomat lopputulokset ovat mahdollisia, kuten että kolikko katoaa heitettäessä esimerkiksi viemäriin, eikä kolikonheiton lopputulosta täten voida lukea. Klassinen todennäköisyys sellaisenaan ei ole usein käytetty tulkinta biologisessa mallintamisessa, mutta se loi tärkeän pohjan frekventistiselle tulkinnalle.

1.3.2 Frekventistinen koulukunta

Frekventistinen koulukunta liittyy läheisesti empirismiin. Klassisessa koulukunnassa todennäköisyys jaetaan tasan kaikkien mahdollisten alkeistapausten kesken *a priori*, kun taas frekventistinen tulkinta tekee päätelmät vasta toistokokeen jälkeen, tulkiten todenäköisyyden toteutuneena suhteellisena frekvenssinä.

Frekventistinen koulukunta siis tulkitsee tapauksen a todennäköisyyden likiarvon olevan

$$p_a \approx \frac{\text{tapausten } a \text{ toteutunut lukumäärä}}{\text{toistojen lukumäärä}}$$

ja tarkan arvon

$$p_a = \lim_{\text{toistot}\to\infty} \frac{\text{tapausten } a \text{ toteutunut lukumäärä}}{\text{toistojen lukumäärä}}.$$

Frekventistisen koulukunnan aksiomaattinen väite on raja-arvon tarkkuuden yhdistäminen elämismaailmamme empiriaan. Frekventistisen koulukunnan kritiikki rakentuukin helposti reaalimaailman pohjalta. Jos kolikkoa ei ole koskaan heitetty, ei sen kruunalle ja klaavalle voi asettaa frekventististä todennäköisyyttä, vaikka intuition mukaan näille tapahtumille selkeästi on olemassa todennäköisyys. Myöskin, jos kolikkoa on heitetty tismalleen kerran, niin toisen puolen frekventistinen todennäköisyys on 1 ja toisen 0. Toisaalta tämä epätarkkuus korjaantuu useammalla toistolla keskeisen raja-arvolauseen ja suurten lukujen lain takia.

Frekventismiä voidaan kritisoida myös tapahtumien yhteydellä toisiinsa. Saman kolikon kaksi peräkkäistä heittoa lienevät riittävän samanlaisia, että mielekkäitä todennäköisyyksiä voidaan laskea usean peräkkäisen heiton sarjasta. Monimutkaiset prosessit ovat kuitenkin usein ainutlaatuisia, kuten esimerkiksi populaatioon geneettisen pullonkaulan aiheuttaneet luonnonilmiöt tai kulttuuritapahtumat kuten eduskuntavaalit. Vaikka tällaisten ilmiöiden pohjalta laskettaisiin frekventistisiä todennäköisyyksiä, niiden hyödyntäminen tulevien tapahtumien mallintamisessa sisältää epätarkkuutta. Frekventistinen koulukunta ei näe välimuotoja tapahtumien välillä. Kuitenkin, jo toteutuneiden tapahtumien *a posteriori* analyysissä, kuten monissa tilastollisissa sovelluksissa frekventistinen koulukunta on hyödyllinen ja laajalti käytetty tulkinta. Frekventistinen koulukunta muodostaa yhdessä uskottavuuskoulukunnan kanssa biologisten prosessien mallintamisessa käytetyimmät todennäköisyystulkinnat.

1.3.3 Uskottavuuskoulukunta

Siinä missä klassinen, frekventistinen ja myöhemmin esiteltävä bayesiläinen koulukunta ovat merkittävässä määrin filosofisia tulkintalähtökohtia todennäköisyyden käsitteelle sinänsä, uskottavuuskoulukunta on suppeampi sovellusohje tilastolliseen päättelyyn. Kuten frekventistinen päättely, myös uskottavuuspäättely vaatii olemassa olevan aineiston, jonka perusteella päättely suoritetaan. Uskottavuuspäättely kuitenkin rajaa sovellusaluettaan vielä enemmän: uskottavuuspäättely tapahtuu osana matemaattista mallinnusta.

Uskottavuustulkinta todennäköisyyksille liittyy läheisesti parametrien estimoimiseen. Uskottavuuspäättelyssä ilmiötä tutkitaan määrittämällä siihen liittyviä tunnuslukuja, eli parametreja. Uskottavuustulkinnan tulkintalähtökohta voidaan ilmaista: "Mallin tuntemattomia parametreja koskeva päättely perustuu ainoastaan mittaustuloksista laskettuun uskottavuusfunktioon. [13]"Uskottavuusfunktion voi yksinkertaistaen määritellä frekventistiseksi todennäköisyydeksi saada toteutuneet mittaustulokset satunnaismuuttujaa tai ilmiötä mitattaessa. Uskottavuusfunktion määrittelyavaruus on tutkittavan parametrin mahdolliset arvot. Uskottavuusfunktion globaali maksimi on *suurimman uskottavuuden estimaatti (SUE)*. Tätä arvoa pidetään kaikkein varmimpana arviona parametrin arvolle, sillä mittaustulokset antavat sille eniten tukea. Frekventististä todennäköisyyslaskentaa käytetään siis epävarmuuden minimoimiseksi menetelmässä, jossa mittaustulokset ja matemaattinen malli yhdessä antavat tietoa tutkittavasta ilmiöstä. Uskottavuuskoulukunta periaatteessa sulkee pois *a priori* määritellyt parametrit malliin, mutta käytännössä mallinnuksessa joudutaan käyttämään reunaehdoista tai muusta teoreettisesta viitekehyksestä postuloituja vakioparametreja, jotta epävarmuus mittauksissa minimoituisi ja tutkimus pystytään rajaamaan kiinnostavimpiin parametreihin. Uskottavuuspäättely on tärkeä osa biologiassa käytettävää mallintamista. Parametreillä itsessään on usein biologinen tulkinta, tai ne antavat epäsuorasti tietoa tutkittavasta ilmiöstä.

1.3.4 Bayesilainen koulukunta

Bayesilainen koulukunta perustuu sekä Bayesin kaavaan, että loogiseen todennäköisyyskäsitteeseen, jota rakensivat mm. W.E. Johnson, J.M. Keynes, H. Jeffreys ja varsinkin Rudolf Carnap osana formaalin, loogisen kielijärjestelmän tutkimuksiaan. Toinen yläkäsite, johon Bayesiläinen koulukunta sisältyy, on subjektiivinen todennäköisyyskäsitys.

Näillä tulkinnoilla on kaksi perustavanlaatuista eroa frekventistiseen ja klassiseen tulkintaan: Alkeistapauksilla ei ole samaa esiintymistodennäköisyyttä, vaan ne pai-

notetaan jollain tavalla. Lisäksi todennäköisyys itsessään käsitetään muuttuvaksi, subjektiiviseksi uskomukseksi tapahtuman esiintymisestä. Todennäköisyyden subjektiivisuutta voidaan käsitellä monella tavalla, mutta tärkeä yhdistävä ohjesääntö on uskomuksien, eli todennäköisyyksien, päivittäminen todistusaineiston valossa.

Käytännössä Bayes-päättelyllä usein viitataan menetelmään, jossa mallin parametreista tehdään päätelmiä ainoastaan niiden *posteriorijakauman* pohjalta. Parametrille oletetaan ennakko-uskomusten perusteella jakauma, jonka muotoja sen luullaan noudattavan. Tähän *priorijakaumaan* yhdistetään soveltuvalla metodilla saadut mittaustulokset, mistä saadaan posteriorijakauma. Ennakko-uskomusta, eli priorijakaumaa on siis päivitetty mittaustuloksilla. Priorijakauman valinta voidaan perustella melko vapaasti, esimerkiksi teoreettisella viitekehyksellä, aikaisemmilla tutkimustuloksilla tai rohkeilla arvauksilla.

Yksinkertainen esimerkki (ajatuskoe) Bayes-päättelystä: Tutkijalla on pohjatietoa, että perinnöllistä sairautta esiintyy n. 12,5 % populaation yksilöistä. Tutkittava populaatio on kuitenkin hyvin pieni, tai suuri osa näytteistä on viallisia: 600 yksilön otoksesta vain neljälläkymmenellä todetaan perinnöllinen sairaus riittävällä varmuudella. Tutkija hyödyntää pohjatietoaan muodostamalla priorin: hän laskee populaatioon mukaan tuhat kuvitteellista yksilöä, joista 125:llä on perinnöllinen sairaus. Tutkija yhdistää todellisen ja kuvitteellisen aineiston posterioriksi, ja laskee sairauden (frekventistisen) esiintyvyyden tämän osin kuvitteellisen posteriorin osalta: $165/1600 \approx 10{,}3\,\%$. Pelkän mittaustuloksen perusteella taudin esiintyvyys olisi ollut $40/600 \approx 6{,}7\,\%$. Tällainen menettely on mahdollisesti hyödyllinen esimerkiksi suuntaa-antavaksi välitulokseksi mittausten edetessä.

1.4 Mallien vertailu

Tutkijoilla on usein intressejä vertailla mallien paremmuutta kvantitatiivisesti. Kysymyksenasettelu voi olla esimerkiksi: "Mikä malli antaa harvinaisille ääritapauksille realistisen painon?", "Kasvaako mallin selitysvoima, jos siihen lisätään parametri?", tai "Mikä täysin erilaisista malleista kuvaa ilmiötä parhaiten?". Näistä kysymyksistä ensimmäinen on erityisen kiinnostava taloustieteen mallintamisessa, mutta sitä ei käsitellä tässä työssä.

Jotta mallien vertailusta voidaan puhua, tarvitaan käsitteitä. Uskottavuusfunktion käsite on tärkeä.

Määritelmä 1.1. Olkoon $p \in \mathbb{R}^n$ mallin parametrivektori. Funktiota

$$\mathcal{L} : \mathbb{R}^n \to \mathbb{R}_+,$$

joka kuvaa parametrijoukon $\pi \in \mathbb{R}^n$ todennäköisyyteen saada mittaustulokset (tai todennäköisyyden tiheysfunktioon jatkuvien mittaustulosten tapauksessa) kutsutaan uskottavuusfunktioksi. Logaritminen uskottavuusfunktio on tämän funktion logaritmi:

$$\mathcal{LL} : \mathbb{R}^n \to \mathbb{R}, \quad \mathcal{LL}(\pi) = \log\left(\mathcal{L}(\pi)\right).$$

Määritelmä 1.2. Parametrivektori $\hat{\pi}$, joka maksimoi funktion $\mathcal{L}$ kutsutaan suurimman uskottavuuden estimaatiksi parametrivektorille p. Funktiota, joka kuvaa mittaustulokset parametrivektoriin $\hat{\pi}$ kutsutaan suurimman uskottavuuden estimaattoriksi.

Tässä tutkielmassa yksiulotteista parametrivektoria merkitään yleensä p ja kutsutaan parametriksi eikä parametrivektoriksi.

1.4.1 Uskottavuuksien suhdetesti

Uskottavuuksien suhdetesti ottaa kantaa kysymykseen: "Kasvaako mallin selitysvoima, jos siihen lisätään parametri?"

Olemassaolevaa mallia laajennetaan lisäämällä siihen parametri p_0 mallintamaan aiemmin huomioimatonta näkökulmaa ilmiöön. Parametri p_0 valitaan siten, että arvolla $p_0 = 0$ laajennettu malli redusoituu laajentamattomaksi, aikaisemmaksi malliksi (p_0 voi olla skalaari tai moniulotteinen parametrivektori). Nimetään laajentamaton malli $M(0)$ ja laajennettu malli $M(p_0)$.

Laajennoksen mielekkyyttä voidaan tutkia uskottavuuden suhdetestillä (*Likelihood Ratio Test* eli LRT). Suhdetesti vaatii kaksi taustaoletusta:

- Mittaustulosten (populaation) koko on riittävän suuri, $N \geq 25$ [6].
- Mittaustulosten sovituksen jälkeen parametri p_0 on normaalijakautunut.

Lisäksi tarpeellinen, mutta usein lausumaton oletus on, että $M(p_0)$ ja $M(0)$ ovat ilmaistavissa yksiselitteisenä tunnuslukuna. Toisin sanoen, malli on funktio, joka tuottaa mittaustuloksista yhden tunnusluvun, joka on vertailukelpoinen toisen mittaustulossarjan tuottaman tunnusluvun kanssa.

Näiden oletusten vallitessa mallien uskottavuuksien suhde $\mathcal{L}(M(p_0))/\mathcal{L}(M(0))$ on χ^2-jakautunut [3].

Määritellään suhteen perusteella logaritminen indeksiluku λ, joka on myös χ^2-jakautunut:

$$\lambda = -2(\mathcal{LL}(M(p_0)) - \mathcal{LL}(M(0))).$$

Vakiotermillä -2 skaalataan mallien uskottavuuksien suhdetta, joka esiintyy tässä indeksiluvun lausekkeessa logaritmisten uskottavuusfunktioiden erotuksena. Lähteestä ja käyttötarkoituksesta riippuen suhde voidaan merkitä myös toisin päin, eli laajentamattoman mallin uskottavuus jaettuna laajennetun mallin uskottavuudella. χ^2-jakauman vapausasteluku k määritellään mallien $M(p_0)$ ja $M(0)$ (skalaari)parametrien määrän erotukseksi. Luottamustasoa ilmaistaan symboleilla γ tai α, joiden määritelmät ovat:

$$0 < \gamma < 0{,}5$$

$$0{,}5 \leq \alpha < 1$$

$$\gamma = 1 - \alpha.$$

Suhdetestin soveltamisessa on kaksi osaa, kummassakin kokeillaan, onko laajennetun mallin hylkäämiselle perusteita.

1) Logaritmisen uskottavuusfunktion "etäisyys nollaparametreista". Määritellään luottamusväli:

$$\mathcal{LL}(\pi_\alpha) = \mathcal{LL}(\pi_\gamma) = \mathcal{LL}(\hat{\pi}) - \chi^2_{k,\gamma}, \quad \pi_\alpha < \pi_\gamma,$$

missä $\hat{\pi}$ on laajennetun mallin suurimman uskottavuuden parametrifunktio. Jos väli $[\mathcal{LL}(\hat{\pi}) - \chi^2_{k,\gamma}, \mathcal{LL}(\hat{\pi})]$ sisältää arvon 0, ei laajentamatonta mallia voida hylätä tilastollisista syistä ja yksinkertaisimman mallin periaatteen mukaisesti laajentamaton malli on riittävä ilmiön kuvaamiseksi.

2) Jos arvo 0 sijoittuu luottamusvälin ulkopuolelle, tarkastellaan indeksiä λ. Jos $\lambda > \chi^2_{k,\gamma}$, niin laajentamaton malli voidaan hylätä valitulla luottamustasolla. Laajennettu malli kuvaa ilmiötä mittaustulosten valossa halutulla luottamustasolla paremmin kuin laajentamaton malli.

1.4.2 Akaiken informaatiokriteeri

Mallien vertailuun voidaan käyttää Akaiken informaatiokriteeriä (AIC) [7]. Akaiken informaatiokriteeri vertaa tutkittavaa mallia hypoteettiseen "oikeaan malliin" Kullback-Liebler-divergenssin avulla. Hirotugu Akaike laati tästä divergenssistä empiirisiin mittaustuloksiin sovellettavan menetelmän. AIC laskee, kuinka paljon informaatiota katoaa mittaustuloksien sovittamisessa malliin. Mallia, jonka sovittaminen säilyttää eniten informaatiota, voidaan pitää AIC:n mukaan parhaana mallina.

Määritelmä 1.3. Akaiken informaatiokriteeri (AIC) määritellään kaavalla

$$AIC = 2\mathcal{LL}(\hat{p}) - 2n_p,$$

missä $\mathcal{LL}(\hat{p})$ on logaritminen uskottavuusfunktio ja n_p mallin parametrien määrä. Pienelle mittauspisteiden lukumäärälle (Vries et al mainitsevat ehdon $N \leq 40$ [7]) määritellään korjattu Akaiken informaatiokriteeri AICc, joka sisältää korjaustermin:

$$AICc = 2\mathcal{LL}(\hat{p}) - 2n_p \frac{N}{N - n_p - 1}.$$

Merkittävästi:

$$\lim_{N \to \infty} AICc = AIC,$$

joten korjatun termin käyttö on perusteltua jokaisessa empiirisessä tapauksessa, missä suurempi laskentatehon tarve ei ole ongelma.

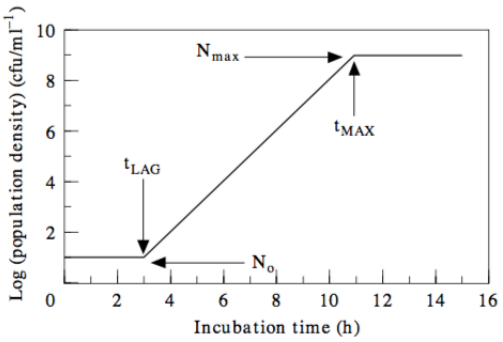
Akaiken informaatiokriteeri toimii, vaikka parametrien määrä vertailtavien mallien välillä olisi erisuuri. Informaatiokriteerin perusteella valitaan malli, jonka AIC tai AICc-arvo on vähiten negatiivinen.

2 Esimerkkejä

2.1 Lineaarinen malli

Joissakin tapauksissa hyvinkin yksinkertainen malli on biologisesti perusteltavissa ja selittää mittaustuloksia kattavasti. Buchanan, Whiting ja Damert vertasivat vuoden 1995 artikkelissaan *When is simple good enough: a comparison of the Gompertz, Baranyi, and three-phase linear models for fitting bacterial growth curves* kolmea eri mallia bakteeripopuulaation kasvun mallintamiseen [5]. Baranyi-malli on differentiaaliyhtälömalli, perustuen bakteeripopulaation kasvua rajoittavien biokemiallisten pullonkaulojen mallintamiseen, kun taas Gompertz-malli on empiirisesti määritettävillä parametreilla skaalattava logistinen käyrä (kts. luku 1.1).

Buchananin, Whitingin ja Damertin artikkelissa käytetty lineaarinen malli on kolmessa palassa määritelty polynomifunktio. Funktion ensimmäinen pala kuvaa sopeutumisvaihetta (*lag phase*) eli aikaa, joka bakteereilla kuluu sopeutua uuteen elinympäristöön. Tutkimuksessa glukoosissa kasvatettuja bakteereja siirrettiin laktoosiin, joten sopeutumisvaiheen aikana bakteerit käynnistivät laktaasientsyymin tuotannon. Sopeutumisvaihe kestää ajanhetkestä $t = 0$ hetkeen $t = t_{LAG}$, joka on merkityksellinen parametri.



Kuva 2: Esimerkki lineaarisesta mallista [5].

Kuvassa 2 logaritmisen Y-akselin yksikkö cfu (*colony forming unit)* tarkoittaa likimain bakteeriyksilöiden määrää. Nimitys tulee mittaustavasta, jossa bakteeriviljelmänäytettä laimennetaan merkittävästi, ja laimennoksesta kasvaneiden uusien kasvustojen lukumäärästä lasketaan näytteen bakteeritiheys.

Eksponentiaalisen kasvun vaihe kestää aikavälin $t_{LAG} < t < t_{MAX}$, jonka aikana bakteerien lukumäärä kasvaa eksponentiaalisesti uuteen ympäristöön siirrettyjen bakteerien lukumäärästä N_0 populaation maksimiarvoon N_{MAX}, jossa ravinnon diffuusionopeus yms. ympäristötekijät degeneroituvat kilpailun vaikutuksesta, aiheuttaen bakteerien jakautumis- ja kuolemisnopeuden tasapainottumisen ja siten populaation kasvun pysähtymisen. Kasvuvaiheen aikana bakteerien lukumäärän logaritmin oletetaan kasvavan lineaarisesti, jolloin bakteerien lukumäärä ajan funktiona on

$$N_t = N_0 + \mu(t - t_{LAG}),$$

missä μ on populaation kasvua kuvaava vakioparametri.

Kolmannessa vaiheessa populaation kasvu on pysähtynyt. Bakteerit ovat joko leporilassa, tai kierrättävät kuolleiden solujen materiaalia samalla nopeudella, kuin uusia bakteereja syntyy jakautumalla.

On merkittävää huomata, että populaatio ei siirry välittömästi vaiheesta toiseen. Bakteerien sopeutumisessa uuteen ympäristöön on pientä varianssia geneettisistä tai biokemiallisista eroista ja/tai vain diffuusion satunnaisuudesta. Toiset bakteerit aloittavat ensimmäisen jakautumisensa aiemmin kuin toiset. Buchanan et al. pohtivat tämän nivelvaiheen mallintamista sigmoidikäyrällä vs. askelfunktiolla, mutta lopulta totesivat tyypillisesti yhden bakteeriviljelmän bakteerien olevan geneettisesti niin samanlaisia, että lineaarinen malli oli riittävän tarkka [5]. Tutkimuksessa vertailtiin lineaarista mallia Gompertz- ja Baranyi-malleihin käyttämällä vertailutunnuslukuna pienintä neliösummaa verrattuna tunnettuun bakteeripopulaatioaineistoon. Lineaarisen mallin huomattiin olevan yksinkertaisuudestaan huolimatta vertailukelpoinen monimutkaisempiin malleihin.

2.2 Hardyn-Weinbergin laki

Kun mendeliläinen genetiikka nousi tiedeyhteisön keskusteluun 1900-luvun alussa, sillä ei osattu selittää genotyyppien jakautumisen jatkuvuutta sukupolvelta toiselle. Dominoivien alleelien väitettiin yleistyvän populaatiossa [39]. G. H. Hardy ja Wilhelm Weinberg keksivät toisistaan erillään suoraviivaisen matemaattisen selityksen alleelien jakautumiselle populaatiossa ilman ulkoisia tekijöitä.

Hardyn-Weinbergin laki on yksi tärkeimmistä modernin populaatiogenetiikan perusteista. Se on esimerkki polynomijakaumasta, tarkemmin trinomijakaumasta. Lain mukaan ilman ulkoisia tekijöitä populaation alleelifrekvenssit pysyvät samana sukupolvesta toiseen, mutta genotyyppien jakauma saavuttaa laskettavissa olevan tasapainoaseman jo yhdessä sukupolvessa.

Yksinkertaisessa esimerkissä geenistä on vain kaksi alleelia: A_1 ja A_2. Merkitään alleelin A_1 frekvenssiä p ja alleelin A_2 frekvenssiä q. Nyt mahdolliset genotyypit ovat A_1A_1, A_1A_2 ja A_2A_2. Olkoon näiden genotyyppien frekvenssit 0,4; 0,4 ja 0,2. Tällöin alleelifrekvenssit ovat $p = 0{,}6$ ja $q = 0{,}4$ [11]. Alleelifrekvenssi on myös todennäköisyys kyseisen alleelin periytymiseen yhdelle jälkeläiselle yhdeltä vanhemmalta. Hardyn ja Weinbergin päättely siis laajentaa Mendeliläistä genetiikkaa todennäköisyyspäättelyllä.

Hardyn-Weinbergin lain merkitys on sen soveltuvuudessa nollahypoteesiksi. Alleelifrekvenssit päätyvät Hardyn-Weinbergin lain mukaiseen tasapainoasemaan jos melko tiukat reunaehdot ovat voimassa (kiinnostavimmat ehdot lihavoitu):

1. Tutkittavat organismit ovat diploidisia
2. Kaikki lisääntyminen on suvullista
3. Sukupolvet eivät ole päällekkäisiä
4. **Parinvalinta on aidosti satunnaista, ts. populaatio on äärettömän suuri**
5. Alleelifrekvenssit ovat identtiset sukupuolten välillä
6. **Mutaatioita, muuttoa populaatioiden välillä tai luonnonvalintaa ei tapahdu**

Hardyn-Weinbergin lain ennustama alleelifrekvenssin tasapainoasema on erityistapaus multinomijakaumasta. Esimerkkitapauksessa kumpikin vanhempi voi edustaa yhtä kolmesta mahdollisesta genotyypistä. Genotyyppi A_1A_1 voi tuottaa vain sukusolun tyyppiä A_1 samoin kuin genotyyppi A_2A_2 voi tuottaa vain sukusolun A_2. Heterotsygootti A_1A_2 sen sijaan voi tuottaa satunnaisesti joko sukusolun A_1 tai A_2.

Taustaoletusten vallitessa jälkeläisen eri genotyyppien todennäköisyydet voidaan laskea joko multinomijakaumalla tai sen erityistapauksella, binomijakaumalla.

Laskun voi myös yksinkertaistaa ja päätellä lopputuloksen muilla keinoin, esimerkiksi perustelemalla tuloksen seuraavan alleelifrekvenssin tulkinnasta todennäköisyytenä kyseisen alleelin periytymiseen jälkeläiselle yhdeltä vanhemmalta. Tällöin saadaan yksinkertainen taulukko (heterotsygootin vanhemmille on kaksi mahdollista kombinaatiota):

$$\begin{aligned} P(A_1A_1) &= p \cdot p = p^2 \\ P(A_1A_2) &= p \cdot q + q \cdot p = 2pq \\ P(A_2A_2) &= q \cdot q = q^2 \end{aligned}$$

Alleelifrekvenssit saadaan myös binomin neliönä:

$$(p+q)^2 = p^2 + 2pq + q^2$$

2.2.1 Hardyn-Weinbergin lain yleistyksiä

Hardyn-Weinbergin lakia voidaan soveltaa alleelifrekvenssin tasapainoaseman ennustamiseen myös tapauksissa, joissa tarkastellaan useampaa kuin kahta geenialleelia.

Kolmen alleelin tapauksessa frekvenssit saadaan trinomin neliöstä:

$$(p+q+r)^2 = p^2 + q^2 + r^2 + 2pq + 2pr + 2qr$$

Ja samoin mielivaltaisen monelle alleelille $(p_1 + \ldots + p_n)^2$ saadaan homotsygooteille A_iA_i ja heterotsygooteille A_iA_j jakaumat:

$$\begin{aligned} P(A_iA_i) &= p_i \cdot p_i = p_i^2 \\ P(A_iA_j) &= p_i \cdot p_j + p_i \cdot p_j = 2p_ip_j \end{aligned}$$

Polyploidisilla organismeilla on useampi kuin kaksi kappaletta kutakin kromosomia. Kromosomikopioiden määrä vaikuttaa polynomin eksponenttiin Hardyn-Weinbergin lain mukaisessa mallinnuksessa. Esimerkkinä tetraploidinen kahden alleelin geeni, jonka alleelifrekvenssiä ennustetaan termillä $(p+q)^4$.

$$
\begin{aligned}
P(A_1A_1A_1A_1) &= p^4 \\
P(A_1A_1A_1A_2) &= 4p^3q \\
P(A_1A_1A_2A_2) &= 6p^2q^2 \\
P(A_1A_2A_2A_2) &= 4pq^3 \\
P(A_2A_2A_2A_2) &= p^4
\end{aligned}
$$

Huomataan, että kyseessä on multinomijakauma, jossa parametri n eli riippumattomien toistokokeiden määrä on organismin ploidialuku, ja k_i kertaa esiintyvät alleelit $A_1 \dots A_n$ puolestaan ovat mahdolliset tapahtumat todennäköisyyksillä $p_1 \dots p_n$. Tällä yleistyksellä voidaan ennustaa yksittäisen mielivaltaisen polyploidisen alleelikombinaation $A*$ frekvenssi tasapainoasemassa multinomijakauman pistetodennäköisyytenä:

$$
P(A*) = \frac{n!}{k_1! \cdots k_n!} p_1^{k_1} \cdots p_n^{k_n}
$$

2.2.2 Hardyn-Weinbergin lain sovelluksia

Hardyn-Weinbergin lain hyödyllisyys on sen taustaoletuksissa, joiden poissaolo voidaan tulkita tutkittavan populaation erityisiksi evoluutiota aiheuttaviksi tekijöiksi. Populaation poikkeamista Hardyn-Weinbergin lain ennustamasta ideaalisesta tasapainoasemasta voidaan tutkia kvalitatiivisten, biologian piirissä olevien päätelmien lisäksi erilaisten matemaattisten tunnuslukujen avulla.

2.2.3 Esimerkki: Kuru-epidemian selviytyjät

Seuraavassa esimerkissä, jonka Freeman ja Herron ja esittelivät oppikirjassaan *Evolutionary Analysis* nähdään, miten yksinkertaisella tunnusluvulla voidaan hylätä nollahypoteesi siitä, että Hardyn-Weinbergn tasapainoasema on voimassa populaatiossa [9].

Hardyn-Weinbergin lain mukainen tasapainoasema ei toteudu populaatiossa, jos alleelifrekvenssit poikkeavat merkittävästi lain ennustamasta tasapainoasemasta. Mead käytti vuonna 2003 Hedrickin aineistoa vuodelta 1908 soveltaakseen χ^2-testiä tasapainoaseman poikkeaman merkittävyyden tarkasteluun [23].

Meadin tutkimuksessa keskityttiin ihmisen prioniproteiinigeenin (PRNP) kodoniin 129. Yksilön genomissa voi olla kyseisen kodonin kohdalla joko metioniini+metioniini (Met/Met), metioniini+valiini (Met/Val) tai valiini+valiini (Val/-

Val). Met/Val heterotsygotian tämän kodonin suhteen arvellaan lisäävän vastustuskykyä prionisairauksille, kuten kurulle.

Tutkimuksessa tarkasteltiin 30 vanhaa Fore-heimon naista Papua-Uudessa-Guineassa, jotka olivat syöneet kuolleita sukulaisiaan, mutta selvinneet kuruepidemiasta sairastumatta. Kulttuurillisista syistä kuolinaterioille osallistuivat pääsääntöisesti heimon naiset.

Heimon osa, joka ei ollut altistunut kuolinaterioille oli tutkittavan geenin suhteen Hardyn-Weinbergin lain ennustamassa tasapainoasemassa.

Tutkittavien henkilöiden genotyypit olivat:

Met/Met	**Met/Val**	**Val/Val**
4	23	3

Freeman ja Harron laskivat tämän aineiston perusteella, oliko kuru-epidemiasta selvinneiden vanhojen naisten populaatio Hardyn-Weinbergin lain mukaisessa tasapainoasemassa.

Otoksessa oli 30 yksilöä, eli 60 kopiota PRNP-geenistä. Siispä alleelifrekvenssit olivat:

$$\text{f(Met)} = p = \frac{(8 + 23)}{60} = 0{,}52$$

ja

$$\text{f(Val)} = q = \frac{23 + 6}{60} = 0{,}48$$

Eri genotyyppien ennustetut tasapainofrekvenssit saadaan Hardyn-Weinbergin laista:

Met/Met	**Met/Val**	**Val/Val**
$p^2 = (0{,}52)^2$	$2pq = 2 \cdot 0{,}52 \cdot 0{,}48$	$q^2 = (0{,}48)^2$
$= 0{,}27$	$= 0{,}50$	$= 0{,}23$

Nämä ennustetut frekvenssit olisivat 30 henkilön populaatiossa:

Met/Met	**Met/Val**	**Val/Val**
$0{,}27 \cdot 30$	$0{,}50 \cdot 30$	$0{,}23 \cdot 30$
$= 8$	$= 15$	$= 7$

Populaatiossa on enemmän heterotsygootteja kuin Hardyn-Weinbergin lain mukainen tasapainoasema ennustaa. Tämä on kiinnostava huomio, ja motivoi tutkimaan tapausta lisää. Joko jokin Hardyn-Weinbergin lain taustaoletuksista ei päde populaatiossa, mikä voi johdattaa populaation erityispiirteiden erittelyyn tutkimuksen

edetessä tai muihin kiinnostaviin biologisiin huomioihin, tai poikkeama ennustetusta tasapainoasemasta on ei-merkittävää seurausta perimän satunnaisuudesta.

Poikkeaman tilastollisen merkittävyyden selvittämiseen on monia työkaluja. Freemanin ja Herronin käyttämä χ^2-testi on yksinkertainen ja nopeasti laskettava testi, joka tässä tapauksessa antaa vahvahkon tuloksen.

χ^2-testissä lasketaan testisuure, jota verrataan χ^2-jakaumaan, mistä nähdään todennäköisyys mittaustulosten saamiseen satunnaisesti nollahypoteesin vallitessa. χ^2-testillä voidaan tarkastella koko monta satunnaisesti periytynyttä geeniä sisältäneen populaation satunnaisen muodostumisen todennäköisyyttä helposti yhdistämällä usea satunnaisesti muodostunut tutkimuskohde yhdeksi summaluvuksi, jonka jakauma tiedetään *a priori*.

$$\chi^2 = \Sigma \frac{(\text{havainnot - ennuste})^2}{\text{ennuste}}$$

$$\chi^2 = \frac{(4-8)^2}{8} + \frac{(23-15)^2}{15} + \frac{(3-7)^2}{7} = 8{,}55$$

Khiin neliö -testin tuloksen tulkintaa varten on pohdittava testin vapausasteluku. Yleisellä tasolla vapausasteluku on havaintoluokkien määrä miinus mittaustuloksista laskettujen tai otettujen, ennustettujen suureiden laskemiseksi välttämättömien erillisten lukujen lukumäärä. Tässä tapauksessa havaintoluokkia on kolme (genotyypit). Hardyn-Weinbergin lain mukaisen tasapainoasema-ennusteen laskemiseksi tarvittiin mittaustuloksista vähintään populaation koko ja toinen alleelifrekvenssi f(Met) tai f(Val). Alleelifrekvenssit ovat toistensa komplementteja, joten ne sisältävät saman informaation. Vapausasteluvuksi saadaan siis $3 - 2 = 1$.

Lukuisien Hardyn-Weinbergin lain sovellusten laskemista varten biologi voi opetella vapausasteluvun laskemiseksi myös yksinkertaisen kaavan tätä erityistapausta varten (Hardyn-Weinbergin lain khiin neliö-testit):

$$\text{vapausasteluku} = \text{genotyyppien lkm} - \text{välttämättömien alleelifrekvenssien määrä} - 1,$$

missä välttämättömien alleelifrekvenssien määrä tarkoittaa vähimmäismäärää välivaiheena laskettavia alleelifrekvensseja, jotta jokaiselle genotyypille voidaan laskea Hardyn-Weinbergin lain mukainen ennustettu yksilömäärä.

Tällä kaavalla saadaan tismalleen sama vapausasteluku: $3 - 1 - 1 = 1$. Khiin neliö on sitä suurempi mitä enemmän havainnot eroavat ennusteesta. Verratessa laskennallista testilukua khiin neliö-jakaumaan, on siis etsittävä pienin mahdollinen

todennäköisyys, jolla khiin neliö on vähintään testiluvun verran, tässä esimerkissä siis se todennäköisyyden P arvo, jolla $\chi^2 \geq 8{,}55$. Taulukoiduista khiin neliö-jakauman arvoista tai itse khiin neliö-jakauman pistetodennäköisyysfunktiosta integroimalla saadaan pienimmäksi mahdolliseksi todennäköisyydeksi $P = 0{,}00346$. Tätä pidetään tilastollisesti tarpeeksi merkittävän pienenä todennäköisyytenä, jotta nollahypoteesi voidaan kyseenalaistaa.

Todennäköiseksi selitykseksi mittaustuloksille siis jää, että PRNP-geeni fore-heimon populaation kuru-epidemiasta selvinneiden vanhojen naisten osapopulaatio ei ole tämän otoksen perusteella Hardyn-Weinbergin lain mukaisessa tasapainoasemassa. Tämä tarkastelu ei kuitenkaan paljasta, mikä Hardyn-Weinbergin lain taustaoletuksista ei pidä paikkaansa. Siihen tarvitaan lisätutkimusta populaatiosta biologian tieteenalan alla, mahdollisesti matemaattisia työkaluja käyttäen.

2.3 Tajiman D-koe

Edellä käsiteltiin Hardyn-Weinbergin lakia esimerkkinä matemaattisesta indikaattorista valintapaineen havaitsemiseksi nollahypoteesi hylkäämällä. Helposti herää kuitenkin kysymys, voitaisiinko monimutkaisemmilla malleilla tai tunnusluvuilla saada enemmän tietoa valintapaineesta, esimerkiksi onko kyseessä suuntaava vai tasapainottava valinta.

Tässä luvussa käsitellään tunnettua tilastollista tunnuslukua geneettisen monimuotoisuuden tarkastelussa: Tajiman D-koetta.

Tajiman D-koe, tai pelkkä Tajiman D itse tunnusluvun nimenä, on tilastollinen testi, joka menee syvemmälle tutkittavan lajin perimään kuin perimää alleelifrekvenssitasolla tarkasteleva Hardyn-Weinbergin laki. Kokeen yleisin käyttötarkoitus on havaita luonnonvalinnan tapahtuminen perimästä. Nollahypoteesinä on, että luonnonvalintaa ei ole tapahtunut, ja tutkittavassa osassa perimää on tapahtunut neutraaleja pistemutaatioita yksilöiden välillä, johtaen eroihin yksilöiden välillä. Jos luonnonvalinta on kohdistunut tutkittavalla alueella sijaitsevaan geeniin, niin usealla yksilöllä on perimässään sama kelpoisuutta (fitness) lisäävä geeni, joilloin niiden perimä on tältä kohdalta samanlainen. Hyödyllisen geenin ympäristössä sijaitsevat emäsparit voivat myös olla samoja yksilöillä, joilla on tämä sama hyödyllinen geeni, sillä ne periytyvät geenin mukana, ja geenin hyödyllisyys suojaa niitä muutokselta. Tätä ilmiötä kutsutaan geneettiseksi liftaamiseksi (genetic hitchhiking). Ajan kuluttua neutraalit pistemutaatiot ja crossing over-ilmiö kuitenkin tuovat varianssia näihin liftanneisiin emäspareihin, joten liftaaminen on voimakkaimmillaan juuri sen jälkeen, kun luonnonvalinta on alkanut vaikuttaa vahvistavasti hyödyllisen geenin frekvenssiin populaatiossa. Tajiman D-koe tarkastelee siis

perimää emäsparitasolla koettaen selvittää onko lajin perimässä yksilöiden välillä tilastollisesti merkittävä määrä vaihtelua, jotta on perusteltua todeta luonnonvalinnan valintapaineen kohdistuneen tarkasteltavaan alueeseen perimässä. Mahdollisia lisäpäätelmiä ovat mm. valintapaineen alkamisen ajankohta sekä millaisen kategorian valintapaine oli kyseessä.

Luonnonvalinnan havaitseminen Tajiman D-kokeella perustuu nollahypoteesin hylkäämiseen: luonnonvalintaa ei todennäköisesti ole kohdistunut tarkasteltavaan perimän alueeseen, jos siinä esiintyy neutraalin molekulaarisen evoluution teorian mukainen tarpeeksi suuri määrä vaihtelua. Tämän teorian esitti Motoo Kimura vuonna 1983 [20]. Teorian mukaan suurin osa mutaatioista on luonnonvalinnan suhteen neutraaleja, ja emäsparitason geneettinen monimuotoisuus saman lajin yksilöiden välillä on suurimmaksi osaksi seurausta neutraaleista mutaatioista ja satunnaisajautumisesta eikä luonnonvalinnasta. Myöhemmin Tomoko Ohta on osoittanut, että mitä suurempi populaation laskennallinen koko on, sitä suuremmalla osalla neutraaleja mutaatioita on merkittävä, vaikkakin vähäinen valintapaino [11] [27]. Mutaation neutraalius on siis kääntäen verrannollinen populaation kokoon, joka on huomioitava Tajiman D-koetta tulkittaessa.

Tarkastelun alla on saman populaation n:ltä yksilöltä otettu DNA-näyte, joista tarkastellaan samaa sijaintia DNA-ketjussa. Tarkasteltava emäspariketju voi olla mielivaltainen tai rajattu esimerkiksi tiettyyn geeniin tai proteiiniin. Saman lajin eri yksilöillä tämä samankohtainen (homologinen) DNA-sekvenssi voi sisältää yhden tai useamman eri emäsparin mutaation, aiheuttaen eroavaisuuksia tarkasteltavissa DNA-sekvensseissä yksilöiden välillä.

Oletuksena vain pienessä osassa emäspareja (nukleotideja) esiintyy varianssia yksilöiden välillä, ainakin jos tutkimusta tehdään yhden lajin sisällä. Luku S on lukumäärä emäsparien paikkanumeroista, joissa varianssia tapahtuu (engl. number of segregating sites). Esimerkiksi jos kaikki eroavaisuudet esiintyvät joko neljännessä tai kahdeksannessa nukleotidissa, olisi $S = 2$. Tämän luvun biologinen tulkinta on lajin genomissa olevien evolutiivisen valintapaineen alaisuudessa olevien kohtien lukumäärä (tutkittavan sekvenssin alueella).

Luku S itsessään ei kuitenkaan ole hyvä tunnusluku lajin DNA:n monimuotoisuudelle ja valintapaineille, sillä se riippuu otoskoosta. Otoskoko tarkoittaa tutkittavien perimänäytteiden lukumäärää ottamatta kantaa niiden pituuteen. Kokeella voidaan tutkia yksittäistä geeniä, osaa siitä tai vaikka koko perimää [34]. Eroavien nukleotidien lukumäärä S jaettuna vertailuparien lukumäärällä $\binom{n}{2}$ sen sijaan on biologisesti merkityksellinen luku, jota merkitään symbolilla π. Yleensä laskettua lukua π kiinnostavampi on sen (mallinnettu) odotusarvo, jota merkitään $E[\pi] = \theta$ tai $E[\pi] = M$, kuten Tajima itse tätä odotusarvoa merkitsee. Itse käytän tar-

kastelussa odotusarvolle symbolia θ. Eri populaatiomalleissa θ on hyvin tärkeä parametri: se mallintaa populaation perinnöllisen muuntelun määrää (laajuutta).

Kun oletetaan populaation olevan tasapainoasemassa siten, että luonnonvalinta ei vaikuta tarkasteltavaan DNA-sekvenssiin, voidaan odotusarvoa θ estimoida kahdella eri tavalla: Tajiman tai Wattersonin estimaattoreilla, joista edellinen perustuu yksilöiden väliseen nukleotidivertailuun ja jälkimmäinen siihen, kuinka monessa paikassa genomissa nukleotidien eroavaisuuksia on koko populaatiossa eli lukuun S.

Tajiman estimaattorissa lasketaan suoraan perimäaineistosta vaihtuvuutta sisältävien nukleotidien lukumäärä suhteutettuna vertailuparien lukumäärään, eli vertaillaan kahta yksilöä pari kerrallaan, jotta saadaan keskiarvo yksilöiden välisten nukleotidierojen lukumäärälle. Tajiman estimaattoria merkitään $\hat{\theta} = \hat{k}$ riippuen lähteestä. Tässä tutkielmassa käytetään merkintää $\hat{\theta}$, koska se on yleisempi moderneissa artikkeleissa. Tästä merkintätavan valinnasta johtuen joissakin kaavoissa esiintyy sekä $\hat{\theta}$ että θ. Näiden symbolien merkitysero on tärkeä tiedostaa Tajiman D-koetta käsitellessä, sillä juuri se on tutkimuskohteena. Tajiman estimaattori määritellään:

$$\hat{\theta} = \frac{\sum_{j=1}^{n} \sum_{i=j}^{n} k_{ij}}{\binom{n}{2}}, \tag{3}$$

missä i ja$j$ ovat eri yksilöiden DNA-näytteiden indeksilukuja aineistossa ja k_{ij} tietyn vertailuparin toisistaan eroavien emäsparien lukumäärä [33].

Toinen estimaattori on ns. Wattersonin estimaattori, jota merkitään lähteestä riippuen $\hat{M}$ tai $\hat{\theta}_W$ [37]. Wattersonin estimaattori määritellään

$$\hat{M} = \hat{\theta}_W = \frac{S}{a_n}, \tag{4}$$

missä a_n on otoskokoa n vastaava harmoninen numero:

$$a_n = \sum_{i=1}^{n-1} \frac{1}{i}.$$

Wattersonin estimaattori perustuu neutraaleihin mutaatioiden/geneettisen ajautumisen (eng. genetic drift) malleihin ja oletukseen populaation Hardyn-Weinbergin tasapainoaseman saavuttamisesta [20] [37]. Eroavien nukleotidien määrän skaalaaminen harmonisella numerolla perustuu populaatiomallien käsitteeseen *varianssin huomioivasta populaation koosta* (eng. effective population size), jossa peräkkäiset,

ei-päällekkäiset sukupolvet palautetaan yhteen numeroon N_e, joka on sukupolvien kokojen harmoninen keskiarvo [17]. Tarkemmin sanottuna N_e olisi se Wright-Fisher -populaatiomallin mukainen populaation koko, jossa geneettinen ajautuminen olisi sama kuin tutkittavalla populaatiolla skaalauksen jälkeen. Tämä harmoninen skaalaus on biologisesti merkittävä, sillä se painottaa populaation pullonkaulojen merkitystä, koska harmoninessa keskiarvossa pienet termit saavat suuren painoarvon.

Wattersonin estimaattori tuo siis biologista tietoa mukaan laskentaan: oletuksen neutraalista geenivirtamallista sekä harmonisen skaalauksen. Populaatiomalleissa geneettisen monimuotoisuuden parametrin θ alkuperäinen määrittelijä onkin juuri Watterson, joka määritteli sen $\theta = 4N_e\mu$ [37], missä N_e on effektiivinen populaatiokoko (kuten ylempänä) ja μ tarkasteltavan lokuksen (paikan kromosomissa) mutaatiotodennäköisyys. Haploideilla olioilla olisi vastaavasti $\theta = 2N_e\mu$, sillä haploideilla olioilla ei ole kahta kromosomia, joista meioosissa satunnaisesti valitaan toinen periytymistä varten. Tämä Wattersonin määritelmä parametrille θ olettaa nollahypoteesin olevan voimassa luonnonvalinnan suhteen: vakiona pysyvä populaation koko, Hardyn-Weinbergin tasapainoasema alleelifrekvenssien suhteen ja tasapaino mutaatioiden sekä geenivirran välillä. Osittain tästä johtuen Tajiman D-koe soveltuu hyvin luonnonvalinnan nollahypoteesin testaukseen.

Tajiman estimaattorin voidaan tulkita tarkoittavan havaittua perimän monimuotoisuutta populaatiossa ja Wattersonin estimaattorin puolestaan odotettua monimuotoisuutta. On huomattavaa, että nämä estimaattorit ovat identtiset, kun otoskoko $n = 2$, joten Tajiman D-koe vaatii vähintään otoskoon $n \geq 3$.

Tajiman D-koe vertaa näitä kahta estimaattoria (Tajiman ja Wattersonin) laskemalla niiden erotuksen jaettuna erotuksen keskihajonnalla. Tajima estimoi erotuksen kovarianssia [34] päätyen esimaattiin

$$\hat{V}(\hat{\theta} - \frac{S}{a_1}) = e_1 S + e_2 S(S-1).$$

Tämän kovarianssin estimaatin ja myöhemmin esitettävän tunnusluvun D parametrit on määritelty liitteessä 1. Tajima perusteli tämän estimaatin aikaisemmassa julkaisussaan [33]. Kyseessä on estimaatti eikä tarkka arvo, sillä varianssin tarkka arvo vaatisi termin $\theta = 4N_e\mu$ tarkan arvon tuntemisen, mikä ei ole realistista todellisilla populaatioilla, joita tutkittaessa N_e on yleensä suuri ja μ estimoitu yleensä suurehkolla epävarmuudella, jos sitä on estimoitu laisinkaan. Estimaattorien kovarianssia ja sen tulkintaa Wright-Fisher populaatiomallin viitekehyksessä on käsitelty myöhemmissäkin tutkimuksissa esimerkiksi pohtimalla varianssin estimaatin pätevyyttä populaatiokoon n eri suuruusluokilla. [18]. $\hat{V}$:n parametrisointi

on esitetty liitteessä 1.

Eri estimaattorien erotuksiin perustuvia, toistensa kaltaisia geneettisen monimuotoisuuden tunnuslukuja voidaan postuloida useita. Tajiman D-koe on kuitenkin simulaatoiden perusteella tämän tunnuslukuperheen vahvimmasta päästä ja se on kirjallisuudessa yleinen koe [30].

Estimaattorien erotuksen estimoidun kovarianssin parametrisoinnin jälkeen voidaan lopulta lausua itse Tajiman D:

$$D = \frac{\hat{\theta} - \frac{S}{a_1}}{\sqrt{\hat{V}(\hat{\theta} - \frac{S}{a_n})}} = \frac{\hat{\theta} - \frac{S}{a_1}}{\sqrt{e_1 S + e_2 S(S-1)}} \tag{5}$$

Varianssin estimoinnin seurauksena Tajiman D:n keskiarvo on noin 0 ja varianssi noin 1. Tajiman D on tilastollinen tunnusluku, jonka biologinen merkitys on tulkittava. D:n tulkintaa varten tarvitaan tietoa D:n jakaumasta.

2.3.1 Tajiman D-kokeen jakauma ja kriittiset arvot

Tajiman D-koetta käytettäessä tulee ensin arvioida, onko saatu D:n arvo tilastollisesti merkittävästi poikkeava nollahypoteesista.

Alkuperäisessä julkaisussaan, jossa D-koe esiteltiin Tajima ehdotti D:n luottamusvälien mallintamista betajakauman avulla, normaalijakaumankin ollessa kelvollinen korkeilla parametrin n arvoilla (tutkittaessa useita emäspareja tai yksilöitä). Tajima vertaili tietokoneella simuloituja D:n arvoja betajakaumaan ja julkaisi D:n luottamusvälijakaumat erilaisilla otoskokoparametrin n arvoilla [34]. Tajima kuitenkin jätti varauksen tulkintaansa: solussa tuman DNA voi uudelleenjärjestyä kopioituessaan (tekijänvaihdunta meioosissa), mikä aiheuttaa erotuksen varianssin olevan estimoitua pienempi, mikä kasvattaa D:n itseisarvoa. Menetelmä siis tuottaa liian varovaisen arvion tunnusluvuksi tutkittaessa tuman DNA:ta. Toisaalta neutraali geenivirran malli sisältää oletuksen, että tekijöidenvaihduntaa ei tapahdu [30] [15], joten betajakauman alakanttiin menevälle arviolle voisi löytä muitakin selityksiä.

Myöhemmät tutkimukset ovat käsitelleet D:n jakaumaa sekä tulkintaa:

King, Wakeley ja Carmi havaitsivat, että Wright-Fisher populaatiomallin alla Tajiman estimaattorin $\hat{\theta}$ varianssi ei lähesty nollaa otoskoon kasvaessa [18], vaan sillä on alaraja:

$$\lim_{n\to\infty} Var\left[\hat{\theta}\right] \gtrapprox \frac{\theta^2}{12N_e} \tag{6}$$

King, Wakeley ja Carm selittivät tätä alarajaa sillä, että mallinnettu olio sukulinjoineen on jonkin tilastollisen populaatiomallin tuote, joka on (satunnaisesti) generoitu mm. parametrin θ itsensä avulla. Mallinnettu yksilö on siis vain yksi erityinen geneettinen kombinaatio, joka olisi voinut syntyä muillakin tavoilla samoista esivanhemmista [18].

Itse pohdin, että jos Kingin, Wakeleyn ja Carmin laskemaan varianssin alarajaan sijoitetaan Wattersonin määritelmä $\theta = 4N_e\mu$ saadaan:

$$\lim_{n\to\infty} Var\left[\hat{\theta}\right] \gtrapprox \frac{4}{3}N_e\mu^2.$$

Tämän esitysmuodon perusteella voidaan huomata termin μ^2 skaalaavan voimakaasti alaspäin effektiivistä populaatiokokoa N_e. On mielenkiintoista, että suurempi mutaatiotodennäköisyys siis pienentää voimakkaasti geneettisen monimuotoisuuden varianssin alarajaa, siis tekee monimuotoisuuden arvioimisesta tarkempaa. Tätä voisi selittää yhteys mutaatiotodennäköisyyden ja mutaatioiden asettumisen (vakiintumisen tai eliminoitumisen) välillä [19]. Merkillepantavaa on myös, että parametrin μ tarkka arviointi on haastavaa. Sen suuruusluokaksi arvioidaan yleensä $10^{-6} - 10^{-7}$ [11] [12] [25].

Kingin, Wakeleyn ja Carmin havainto onkin, että Tajiman D-kokeen tarkennukseksi tutkijan kannattaa kasvattaa tutkittavien lokusten laajuutta ennemmin kuin otoskokoa yhtä lokusta kohti (yksilömäärää) [18]. Tämä ohje on tärkeä tutkittaessa populaation tai populaatioiden geneettistä monimuotoisuutta, mutta Tajiman D-koetta voidaan soveltaa myös muihin tarkoituksiin, kuten geneettisten pullonkaulojen, suuntaavan vallinnan tai populaation alipopulaatioihin jakautumisen tunnistamiseksi (lajiutumisen alkaminen), mihin on löydetty muita kalibrointiohjeita. Simonsen et al. muun muassa huomasivat, että geneettistä liftaamista testattaessa suurempi otoskoko vahvisti D-kokeen erottelukykyä enemmän kuin tutkittavien lokusten laajentaminen [30]. Biologiselta kannalta tutkittavan lokuksen laajuuden kasvattaminen vähentää liftanneiden emäsparien sisäistä vaihtelua crossing over-ilmiön takia (laajemmalla alueella on suurempi todennäköisyys sisältää crossing over-emäspareja, jotka ovat periytyneet samankaltaisina). Tämän eron suuruus crossing over-emäsparien ja geneettisen satunnaisajautumisen välillä liftanneiden emäsparien vaihtelua tarkastellessa lienee kuitenkin pieni, sillä satunnaisajautuminen peittää neutraalien crossing-over geenien vähäisen vaihtelun ajan kuluessa.

Lisäksi crossing-overin todennäköisyys on verrannollinen tutkittavan kodonin etäisyyteen kromosomin sentromeeristä. Tämä mahdollinen virhelähde on kuitenkin hyvä huomioida, jos D-kokeella havaitaan heikkoa tukea nollahypoteesin hylkäämiselle lähimenneisyydessä tapahtuneen luonnonvalinnan puolesta.

Sekä k että S ovat luonnollisia lukuja, joten niiden (erotuksen) jakauma on diskreetti. Tämän vuoksi Tajiman D:n jakaumalle asetetut luottamusvälit eivät saavuta tismalleen haluttuja mielivaltaisia luottamustasoja α_n (esim. 95% luottamusväliä).

Simonsen, Churchill ja Aquardo julkaisivat tutkimuksen geneettisen monimuotoisuuden tilastollisten testien nollahypoteesista vuonna 1995 [30]. Artikkelissa vertailtiin Tajiman D-koetta sekä samankaltaista testiä, jonka Fu ja Li kehittivät vuonna 1993 [10]. Näiden erilaisista erotuksista muodostuvien tunnuslukujen odotusarvon tulisi olla nolla, kuvastaen siis nollahypoteesina valintapaineetonta tilannetta neutraalin geenivirtamallin alla. Simulaatiot kuitenkin osoittivat, että odotusarvo on hieman negatiivinen [30]. Tajiman D-kokeen tapauksessa tämä tarkoittaa sitä, että Wattersonin estimaattori on hieman Tajiman estimaattoria voimakkaampi ja D-kokeen jakauma implisiittisesti vasemmalle vino. Simonsen et al. selittivät tämän johtuvan varianssin estimoinnista: tunnuslukujen laskennassa murtolausekkeen nimittäjä riippuu mittaustuloksista. Lisäksi, koska nimittäjä saattaa olla nolla, Simonsen et al. ehdottavat tunnusluvun määrittämistä nollaksi, kun $S = 0$. Tällä on merkittävä seuraus: nollahypoteesia ei voida sulkea pois millään tarkastelluista tunnusluvuista jos geneettistä muuntelua ei havaita [30]. Käytännön tutkimuksessa otoskokoa (yksilömäärä tai tarkasteltavien genomisekvenssien määrä/laajuus) kasvattamalla muuntelun havaitsemattomuuden todennäköisyys voidaan minimoida.

Tajima laski kriittisiä arvoja D-kokeelleen olettamalla D:n noudattavan *suunnilleen* beta-jakaumaa odotusarvolla nolla ja varianssilla 1 skaalattuna mittaustuloksien perusteella lasketulle välille $[D_{min}, D_{max}]$. Tajima perusteli tämän silmämääräisellä yhdenmuotoisuudella beta-jakauman ja D-kokeen simulaatioiden välillä [34]. Kuten aiemmin todettiin, Tajiman menetelmällä D-kokeen tulkinta on liian konservatiivista, mikä heikentää menetelmän tehoa hylätä nollahypoteesi.

Tunnusluvun D jakauman mallintamista hankaloittaa tuntematon parametri θ. Jos parametrin θ parametriavaruus on rajoittamaton, ja sen supremum $\sup_\theta$ lasketaan numeerisesti, niin lasketun maksimin globaalius voi olla kyseenalaista [2]. Parametriavaruutta voi rajoittaa aineiston perusteella erilaisten estimaattien $\hat{\theta}$ avulla, mutta rajoittamattomallekin tapaukselle löytyy laskennallisesti keveä ratkaisu. Lisäksi estimoinnin avulla konstruoidut D-kokeen kriittiset arvot eivät välttämättä ole päteviä [32].

Simonsen et al. sovelsivat menetelmää, jonka Berger & Boos kehittivät vuonna 1992 [2] [30]. Menetelmän käytön tavoitteena on laatia D-kokeelle kriittisiä arvoja eri parametrin S arvoile siten, että kriittiset arvot vastaavat haluttua luottamusastetta α (esim. $\alpha = 0{,}05$). Lisäksi tavoitteena on hylätä nollahypoteesi jos ja vain jos mittaustuloksia ei voida selittää millään positiivisella parametrin θ arvolla (siis millään arvolla, koska parametrin θ määrittelystä seuraa sen positiivisuus). Ehdoista saadaan muodostettua ehto kriittiselle, tiettyä luottamusastetta *alpha* vastaavalle kriittiselle arvovälille $[D_{ala},D_{ylä}]$:

$$\sup_{\theta\in[0,\infty)} [P_\theta(D \leq D_{ala}) + P_\theta(D \geq D_{ylä})] \leq \alpha \tag{7}$$

Simonsen et al. sovelsivat Bergerin ja Boosin menetelmää Tajiman D-kokeeseen valitsemalla pienen luvun β siten, että $\beta < \alpha$ (esim. $\alpha = 0{,}05$ ja $\beta = 0{,}01$). Mittaustuloksien perusteella estimoitiin luottamusväli C_β, joka on $1-\beta$ luottamusväli parametrille θ. Tämän apuluottamusvälin avulla saatiin päivitettyä kriittinen arvoväli $[D_{ala},D_{ylä}]$ laskennallisesti huomattavasti helpompaan muotoon:

$$\sup_{\theta\in C_\beta} [P_\theta(D \leq D_{ala}) + P_\theta(D \geq D_{ylä})] \leq \alpha - \beta \tag{8}$$

Parametriavaruus C_β voidaan nyt jakaa halutulla tavalla laskennallisiin alkioihin θ_i, jotka peittävät koko parametriavaruuden C_β. Jokaiselle arvolle θ_i lasketaan luottamusastetta $\alpha - \beta$ vastaavat kriittiset arvot $[D^\theta_{ala},D^\theta_{ylä}]$. Näistä arvoista valitaan pienin ja suurin arvo vastaamaan alkuperäistä luottamusastetta α:

$$D_{ala} = \min_{\theta\in C_\beta} D^\theta_{ala}, \quad D_{ylä} = \max_{\theta\in C_\beta} D^\theta_{ylä}$$

Tätä kiertotietä käyttäen konstruoidut kriittiset arvot tuottavat todellakin yhtäpitävästi luottamustason α testin [2]. Vielä on kuitenkin konstruoitava tämä $1-\beta$ luottamusväli jokaista θ_i varten. Simonsen et al. sovelsivat tähän Tavarén vuonna 1984 määrittelemää parametrin S jakaumaa, kun θ on tunnettu (lisäksi n tunnetaan ja s määritellään tietyksi havainnoksi satunnaismuuttujana pidetystä parametrista S) [35]. Tavarén määrittelemä kertymäfunktio on:

$$F(s,n,\theta) = P(S \leq s|\theta) = 1 - \sum_{r=1}^{n-1} (-1)^{r-1} \binom{n-1}{r} \left(\frac{\theta}{r+\theta}\right)^{s+1} \tag{9}$$

On huomattava, että S on diskreetti satunnaismuuttuja ja sen kertymäfunktiotakin käsitellään diskreettinä. Tarvittaessa portaattoman approksimaation voi konstruoida laskentaa varten esimerkiksi jos halutaan saada täsmälleen tietty parametriavaruus C_β. Tämän kertymäfunktion avulla voidaan määrätä parametriavaruus C_β kahdella yhtälöllä:

$$P(S \geq s \,|\theta = \theta_{ala}) = \beta/2 \tag{10}$$

$$P(S \leq s \,|\theta = \theta_{ylä}) = \beta/2 \tag{11}$$

Kirjoitetaan yhtälöt 10 ja 11 kertymäfunktion 9 avulla:

$$F(s-1,\, n,\, \theta_{ala}) = 1 - \beta/2 \tag{12}$$

$$F(s,\, n,\, \theta_{ylä}) = \beta/2 \tag{13}$$

Yhtälöt 12 ja 13 ovat numeerisesti ratkaistavissa parametrien θ_{ala} ja $\theta_{ylä}$ suhteen. Simonsen et al. ehdottavat tapauksen $S = 0$ ratkaisemista asettamalla $\theta_{ala} = 0$ ja käyttämällä yhtälön 13 sijasta yhtälöä

$$F(0,\, n,\, \theta_{ylä}) = \beta \tag{14}$$

Simonsen et al. kuitenkin painottavat valinneensa tapauksen $S = 0$ johtavan itse tunnusluvun D määrittelyyn $D = 0$, jolloin kriittisten arvojen tarkastelu ohitetaan ja siirrytään suoraan tunnusluvun biologiseen tulkintaan [30]. Jos tällainen tapaus tapahtuu todellisessa tutkimuksessa esimerkiksi hyvin pienen otoskoon vuoksi, on hyvä pohtia onko tutkimus mahdollista toistaa suuremmalla otoskoolla tai laajemmalla tutkimuksen kohteena olevalla genomin osalla. Myös itse Tajiman D-kokeen käyttö on kyseenalaista tällaisessa tapauksessa.

Todellisessa tutkimustilanteessa voi riittää yksittäisen parametriparin $(n,\, S)$ muodostavat kriittiset arvot luottamustasolle α. Tutkittaessa populaatiomalleja kuitenkin joudutaan usein tilanteeseen, jossa vertaillaan mallin käyttäytymistä eri parametreilla esimerkiksi Tajiman D-kokeen avulla. Tarve useiden eri kriittisten lukujen määrittämiseen voi syntyä esimerkiksi myös mallinnettaessa tuntematonta populaatiota, josta on vain yksittäisiä geneettisiä otoksia, joiden avulla estimoidaan populaation parametreja kuten parametria θ tai $\frac{\theta}{\mu} = 4N_e$.

Simonsen et al. eivät tutkineet parametrin β valinnan vaikutusta kriittisiin arvoihin, mutta kommentoidessaan simulaatioiden virhearviointia he totesivat parametrin β aiheuttavan enintään β suuruisen nollahypoteesin väärän hylkäämisen mahdollisuuden, joka tosin sisältyy luottamustasoon α.

2.4 Tajiman D-kokeen biologinen tulkinta

Tajiman D-kokeen ehkä suurin hyöty on populaatiogeneettisten mallien vertailussa ja nollahypoteesin hylkäämisessä (luonnonvalinnan toteamisessa) kriittisten arvojen avulla. Tunnusluvulle D voidaan kuitenkin antaa biologinen, laadullinen tulkinta. D-koe käsittelee geneettistä monimuotoisuutta, joten on luontevaa tulkita erilaisten D:n arvojen antavan tukea erilaisille tavoille määritellä perimän heterogeenisuus populaatiossa.

Populaation geeniperimän monipuolisuus tietyllä ajanhetkellä on mitattavissa. Tajiman D-kokeen näkökulmasta onkin usein kiinnostavampaa tutkia, miten nykyhetken tilanteeseen on päädytty. Luonnon tapahtumat, kuten olosuhteiden muuttumisen aiheuttama valintapaine, muuttavat eri geenialleelien frekvenssejä populaatiossa sen sopeutuessa muutoksiin.

Tajiman D:n arvosta voidaan tulkinnallisesti havaita mm. tasapainottava luonnonvalinta, positiivisen luonnonvalinnan aiheuttama alleelin fiksaatio ja populaation koon muutokset [8]. Alleelin fiksaatio tarkoittaa sen leviämistä koko populaatioon, saavuttaen frekvenssin 1. Jokaisen alleelin frekvenssi saavuttaa ajan kuluessa populaatiogeneettisten mekanismien vaikutuksesta joko frekvenssin 0 tai 1 [19]. Tasapainottava valinta puolestaan on luonnonvalinnan muoto, joka ohjaa populaatiota monimuotoisuutta kohti esimerkiksi tilanteessa, jossa heterotsygooteilla on suurempi kelpoisuus kuin homotsygooteilla [11].

Muutos elinympäristössä voi aiheuttaa tarkasteltavan geenin fiksaation hyvinkin nopeasti, jos siihen kohdistuu voimakas valintapaine, eli suuntaava luonnonvalinta [11]. On huomattava, että fiksaatio voi tapahtua myös geneettisen satunnaisajautumisen vaikutuksesta. Pienillä populaatioilla tämä on jopa todennäköisempää kuin luonnonvalinnan aiheuttama fiksaatio. Suurilla populaatiokoilla taas heikompi luonnonvalinta riittää aiheuttamaan fiksaation ilman että satunnaisajautuminen sumentaa sen vaikutuksen havaitsemattomiin [19]. Luonnonvalinnan aiheuttamaa fiksaatiota kutsutaan valintapyyhkäisyksi (*selective sweep*). Jos Tajiman D-kokeella tutkitaan aluetta genomissa, joka on hiljattain kokenut valintapyyhkäisyn, voi D:n odottaa olevan negatiivinen: Valintapyyhkäisyn jälkeen suurin osa alleeleista, ellei jopa kaikki, tietyssä genomin lokuksessa ovat samoja. Tällöin uudet mutaatiot ovat automaattisesti harvinaisia, sillä niitä joko ei ole populaatiossa ennestään ollenkaan, tai niitä esiintyy osalla siitä populaation osasta, johon valintapyyhkäisy ei vaikuttanut. Tilanteessa, jossa genomissa esiintyvä vaihtelu koostuu pääosin harvinaisista mutaatioista, Tajiman estimaattori $\hat{\theta}$ aliarvioi geneettistä monimuotoisuutta verrattuna Wattersonin estimaattoriin $\frac{S}{a_n}$, mikä aiheuttaa tilanteen $D < 0$ tai tarkemmin: tunnusluvun D absoluuttinen arvo on enemmän negatiivinen kuin valittu kriittinen arvo. Negatiivisen D:n havaitseminen voidaan

siis tulkita todistusaineistoksi valintapyyhkäisystä ja siten luonnonvalinnasta. Tämä päätelmä on sitä perustellumpi, mitä suurempi populaatio on kyseessä.

Tasapainottavan valinnan ollessa voimassa populaatiossa tutkittavan lokuksen yleisimpien alleelien frekvenssit ovat samankaltaisia. Tällöin lokuksessa on tietty, suuri määrä emäspareja, joissa esiintyy variaatiota (kilpailevat alleelit), ja tämän joukon sisällä on runsaastio variaatiota: Esimerkiksi tilanteessa, jossa kilpailevat alleelit esiintyvät suhteessa 1:1, geeni on kahdentoista emäsparin mittainen ja alleelit eroavat toisistaan kolmen emäsparin suhteen. Tällöin variaatiota löytyisi kolmen emäsparin kohdalla ja joka toisella yksilöllä nämä kolme emäsparia olisivat erilaiset. Tällöin jokaista kolmea monimuotoisuutta sisältävää genomin kohtaa vastaisi $3\sum_{i=1}^{n/2} = \frac{3(n+1)^2-3}{8}$ yksittäistä emäsparieroa. Tämän seurauksena Tajiman estimaattori on suurempi kuin Wattersonin estimaattori, jolloin tunnusluku D saa positiivisen arvon. Mitä suurempi populaatio, sitä vahvempi tämä tulkinta on. Mitä pienempi populaatio ja suurempi mutaatiotodennäköisyys, sitä isommalla painolla positiivinen D on seurausta (neutraaleista) mutaatioista eikä luonnonvalinnasta. Neutraalit mutaatiot kasvattavat tunnusluvun D absoluuttista arvoa tasapainottavan luonnonvalinnan ylläpitäessä variaatiota sisältävien genomin kohtien korkeaa lukumäärää. Tällöin mutaatiot saattavat nostaa D:n arvon jonkin kriittisen arvon yläpuolelle, jolloin satunnaisuutta kontrolloiva kriittinen arvo ylitetään juurikin satunnaismuuttujan realisoitumisen takia.

Luonnonvalinnan toteaminen heikosta Tajiman D-kokeen tuloksesta on riskialtista. Vaikka tilastollinen mielekkyys saavutetaankin kriittisiä arvoja käyttämällä, geneettisen satunnaisajautumisen aiheuttama luonnonvalintaa sumentava vaikutus heijastuu tunnusluvun D arvoon eri mekanismeilla erilaisissa populaation tilanteissa. Varsinkin pieniä populaatioita tutkittaessa tulee geneettisen satunnaisajautumisen voimakkuuden takia olla erityisen varovainen tehtäessä tulkintoja pelkästään Tajiman D-kokeen perusteella [19] [27].

Populaation koon muutoksella on suuri vaikutus Tajiman D-kokeen tuloksiin. Onkin luontevaa käyttää D-koetta havaitsemaan populaatiokoon muutoksia jonkin koko populaatiota kohdanneen ilmiön seurauksena. Geneettiseksi pullonkaulaksi kutsutaan ilmiötä, jossa populaation koko supistumisen jälkeen laajenee. Pullonkaula voi johtua esimerkiksi uudesta ympäristötekijästä, joka heikentää tiettyä alleelia kantavien yksilöiden lisääntymistehoa. Kyseessä voi olla myös äkillinen katastrofi, esimerkiksi ihmisen aiheuttama elinympäristön muutos, joka tappaa yksilöitä alleeleista riippumatta. Kun populaatio kutistuu satunnaisesti kuolemalla pullonkaulan aikana, populaatiossa laajalti levinneitä alleeleja kantavat yksilöt kuolevat todennäköisemmin, koska niitä on enemmän. Tämä muistuttaa heikkoa tasapainottavaa luonnonvalintaa. Tunnusluvun D absoluuttinen arvo kasvaa pullonkaulan aikana. Selviytymistä vähentäneen tekijän poistuttua eliöiden lisääntymisen odo-

tusarvo kasvaa, jolloin mutantteja, heterotsygootteja ja aiemmin ei-optimaalisten alleelien suhteen homotsygootteja lisääntyy enemmän kuin pullonkaulan aikana. Tämä aiheuttaa D:n arvon äkillisen romahtamisen negatiiviseksi. Jos pullonkaulan aiheuttama kuolleisuus on hyvin suuri, aiheuttaen miltei sukupuuton, on se todennäköisesti aiheuttanut myös lukuisten geenien fiksaation tai lähes fiksaation. Tällöin välittömästi pullonkaulan jälkeen minimin saavutettuaan tunnusluku D kasvaa nopeasti lukuisten uusien pistemutaatioiden aiheuttaessa uusia varianssia sisältäviä kohtia genomissa, minkä jälkeen muutosnopeus pienenee [8]. Vähemmän äärimmäisen kuolleisuuden aiheuttaneen pullonkaulan jälkeen tunnusluvun D kasvunopeus on romahduksen jälkeen maltillinen heterotsygotian yleistyessä ja harvinaisten alleelien muuttuessa hieman yleisemmiksi. Ajan kuluessa D lähestyy arvoa 0 pullonkaulan jälkeen, muttei saavuta sitä, sillä D:n "tasapainoasema" on aina hieman negatiivinen geneettisen satunnaisajautumisen, pistemutaatioiden ja fiksaation takia: vaikka monimuotoisuutta sisältäviä genomin kohtia poistuu fiksaation myötä, pistemutaatiot synnyttävät uusia kohtia yhtä suurella nopeudella [20] [19]. Tajiman D-kokeella voidaan siis havaita myös populaation kokema geneettinen pullonkaula, varsinkin jos on mahdollista tehdä seurantatutkimusta useamman sukupolven kohdalla. Pullonkaulan ja luonnonvalinnan signaalit ovat kuitenkin samankaltaisia, ja niiden erottamisessa toisistaan on hyödyllistä käyttää muita biologisia havaintoja, varsinkin jos D:n absoluuttinen arvo on pieni.

3 Lähdeluettelo

[1] Baranyi, J & Roberts, T.A. *Mathematics of predictive food microbiology.* International Journal of Food Microbiology. 26 (1995) 199-218.

[2] Berger, R. & Boos, D. (1992). *P-values Maximized Over a Confidence Set For a Nuisance Parameter.* Journal of the American Statistical Association. 89:1012-1016.

[3] Berger, R., Casella, G. ja Berger, R.L. *Statistical Interference.* Duxbury Resource Center, Pacific Grove, CA, 2001.

[4] Braun, M. (1983) *Differential Equations and Their Applications (Third Edition) Short Version.* Springer-Verlag New York Inc. New York.

[5] Buchanan, Whiting & Damert (1997) *When is simple good enough: a comparison of the Gompertz, Baranyi, and three-phase linear models for fitting bacterial growth curves* Food Microbiology, Volume 14, Issue 4. https://doi.org/10.1006/fmic.1997.0125

[6] Causton, D.R. (1983) *A Biologist's Basic Mathematics.* Edward Arnold Publishers Ltd. Baltimore.

[7] De Vries, G. et al. (2006) *A Course in Mathematical Biology: Quantitative Modeling with Mathematical and Computational Methods.*

[8] Durvasula, A. (2015). *Interpreting Tajima's D.* https://arundurvasula.wordpress.com/2015/02/18/interpreting-tajimas-d/comment-page-1/ Luettu 16.1.2019.

[9] Freeman, S. & Herron, J.C. (2007) *Evolutionary Analysis, Fourth Edition* Pearson. New Jersey.

[10] Fu, Y. & Li, W. (1993). *Statistical Tests of Neutrality of Mutations.* Genetics. 133(3):693-709.

[11] Futuyma, D. (2013) *Evolution, Third Edition* Sinauer Associates Inc. Massachusetts.

[12] Haag-Liutard, C., Dorris, M., Maside X Macaskill, S. et al. (2007). *Direct estimation of per nucleotide and genomic deleterious mutation rates in Drosophilia.* Nature. 445:82-85.

[13] Hájek, A. (2012) *Interpretations of Probability.* Stanford Encyclopedia of Philosophy (Winter 2012 edition), toimittanut Zalta, E. N. Arkistoitu internet-

osoite: https://plato.stanford.edu/archives/win2012/entries/probability-interpret/

[14] Hardy, G. H. (1908) *Mendelian proportions in a mixed population.* Science 28:49-50.

[15] Hudson, R.R. (1990). *Gene genealogies and the coacescent process.* Oxford Surveys in Evolutionary Biology. 7:1-44.

[16] Kalbfleisch, J.G. (1985) *Probability and Statistical Inference, volume 2: Statistical Inference, Second Edition.* Springer. New York.

[17] Karlin, S. (1968). *Rates of Approach to Homozygosity for Finite Stochastic Models with Variable Population Size.* The American Naturalist. 102 (927): 443-455.

[18] King, L., Wakeley, J. & Carmi, S. (2018). *A non-zero variance of Tajima's estimator for two sequences even for infinitely many unlinked loci.* Theoretical Population Biology. 122: 22-29.

[19] Kimura, M. & Ohta, T. (1968). *The average number of generations until fixation of a mutant gene in a finite population.* Genetics. 61:763-771.

[20] Kimura, M. (1983). *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Lontoo.

[21] Laplace, P. S (1814) *A Philosophical Essay on Probabilities.* English edition (published 1951), Dover Publications Inc. New York.

[22] MacKay, D. J. C. (2002) *Belgian euro coins: 140 heads in 250 tosses - suspicious?* University of Cambridge, Department of Physics. http://ww.inference.phy.cam.ac.uk/mackay/abstracts/euro.html, luettu 29.3.2019.

[23] Mead, S., M. P. H. Stumpf, et al. (2003) *Balancing selection at the prion protein gene consistent with prehistoric kurulike epidemics.* Science 300:640-643.

[24] Migon, H.S, Gamerman, D. & Louzada, F. (2014) *Statistical Inference: An Integrated Approach, Second Edition.* Chapman and Hall/CRC. Boca Raton.

[25] Nachman, M.W. & Crowell S.L. (2000). *Estimate of the Mutation Rate per Nucleotide in Humans.* Genetics. 156:297-304.

[26] Newman, James R. (1956) *The World of Mathematics.* Simon & Schuster. New York.

[27] Ohta, T. (2002). *Near-neutrality in evolution of genes and gene regulation.* PNAS. 99 (25): 16134-16137.

[28] Pawitan, Y. (2001) *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press. Oxford.

[29] Russell, B. (1927) *An Outline of Philosophy*. Allen & Unwin. Lontoo, 1927.

[30] Simonsen, K.L., Churchill, G.A. & Aquardo, C.F. (1995). *Properties of statistical tests of neutrality for DNA polymorphism data.* Genetics. 141(1):413-429.

[31] Sprott, D. A. (2000) *Statistical Inference in Science*. Springer. New York.

[32] Storer, B.E. & Kim, C. (1990). *Exact Properties of Some Exact Test Statistics for Comparing Two Binomial Proportions.* Journal of the American Statistical Association. 85:146-155

[33] Tajima, F. (1983). *Evolutionary relationship of DNA sequences in finite populations.* Genetics. 105: 437-460.

[34] Tajima, F. (1989). *Statistical method for testing the neutral mutation hypothesis by DNA polymorphism.* Genetics. 123 (3): 585–95.

[35] Tavaré, S. (1984). *Line-of-descent and genealogical processes and their applications in population genetics models.* Theoretical Population Biology. 26:119-164.

[36] Watanabe, S. (1969) *Knowing and guessing: a quantitative study of inference and information*. Wiley. New York.

[37] Watterson, G.A., (1975). *On the number of segregating sites in genetical models without recombination.* Theoretical Population Biology 7 (2): 256–276.

[38] Wittgenstein, L. (1921) *Tractatus Logico-Philosophicus*. Annalen der Naturphilosophie. 14:191-262.

[39] Yule, G. U. (1902) *Mendel's Laws And Their Probable Relations To Intra-Racial Heredity.* New Phytologist, 1: 222-238, 1902. doi:10.1111/j.1469-8137.1902.tb07336.x

Liite 1 Tajiman D-kokeen parametrit

Tajiman D-kokeessa geneettisen monimuotoisuuden estimaattorien $\hat{\theta}$ ja $\hat{M}$ erotuksen kovarianssin estimoinnin esitys ja parametrisointi Tajiman [34] mukaan:

$$\hat{V}(\hat{\theta} - \frac{S}{a_1}) = e_1 S + e_2 S(S-1),$$

missä

$$e_1 = \frac{c_1}{a_1},$$

$$e_2 = \frac{c_2}{a_1^2 + a_2},$$

$$c_1 = b_1 - \frac{1}{a_1},$$

$$c_2 = b_2 - \frac{n+2}{a_1 \cdot n} + \frac{a_2}{a_1^2},$$

$$b_1 = \frac{n+1}{3(n-1)},$$

$$b_2 = \frac{2(n^2+n+3)}{9n(n-1)},$$

$$a_1 = \sum_{i=1}^{n-1} \frac{1}{i},$$

ja

$$a_2 = \sum_{i=1}^{n-1} \frac{1}{i^2}.$$